Planner: Cost-efficient Execution Plans Placement for Uniform Stream Analytics on Edge and Cloud

Laurent Prosperi
ENS Paris-Saclay, Inria
Cachan, Rennes (France)
laurent.prosperi@ens-paris-saclay.fr

Alexandru Costan, Pedro Silva, Gabriel Antoniu
Univ Rennes, Inria, CNRS, IRISA
Rennes (France)
alexandru.costan@irisa.fr, pedro.silva@irisa.fr, gabriel.antoniu@inria.fr

***Abstract*—Stream processing applications handle unbounded and continuous flows of data items which are generated from multiple geographically distributed sources. Two approaches are commonly used for processing: Cloud-based analytics and Edge analytics. The first one routes the whole data set to the Cloud, incurring significant costs and late results from the high latency networks that are traversed. The latter can give timely results but forces users to manually define which part of the computation should be executed on Edge and to interconnect it with the remaining part executed in the Cloud, leading to sub-optimal placements. In this paper, we introduce Planner, a middleware for uniform and transparent stream processing across Edge and Cloud. Planner automatically selects which parts of the execution graph will be executed at the Edge in order to minimize the network cost. Real-world micro-benchmarks show that Planner reduces the network usage by 40% and the makespan (end-to-end processing time) by 15% compared to state-of-the-art.**

***Index Terms*—stream processing, Edge analytics, hybrid stream processing**

I. Introduction

The age of offline-only Big Data analytics is over, leaving room to online and interactive processing. The proliferation of small sensors and devices that are capable of generating valuable information in the context of the Internet of Things (IoT) has exacerbated the amount of data flowing from all connected objects to private and public cloud infrastructures. The applications leveraging these data (e.g. monitoring, video streaming) raise specific challenges, as they typically have to handle small data (in the order of bytes and kilobytes), arriving at high rates, from many geographical distributed sources and in heterogeneous formats, that need to be processed and acted upon with high reactivity in near real-time.

Two axes are currently explored *separately* to achieve these goals: *Cloud-based analytics* and *Edge analytics*. The traditional approach of sending the data from (potentially millions of) Edge devices to the Cloud for processing was largely adopted due to simplicity of use and the perception of unlimited resources. Here, a plethora of stream processing engines - SPEs (like Spark [34], Flink [19], Kafka [27], Storm [1], Samza [2], Pulsar [3]) are used for data analytics and persistence. In this case, the Edge devices are used just as proxies only to forward data to the Cloud. However, pushing all the streams to the Cloud incurs significant latencies as the wide area networks that are traversed have limited available bandwidth. On the other hand, since Edge devices are getting more powerful and energy-efficient, another vision is to perform an important part of the analysis at the collection site, at the edge of the network. Such an approach allows to take local decisions and enables the real-time promise of the analytics, improving the reactivity and "freshness" of the results. Several Edge analytics engines emerged lately (e.g. Apache Edgent [4], Apache Minifi [5]) enabling basic local stream processing on low performance IoT devices.

More recently, a new *hybrid* approach tries to *combine both Cloud and Edge analytics* in order to offer better performance, flexibility and monetary costs for stream processing. First, processing live data sources can offer a potential solution that deals with the explosion of data sizes, as the data is filtered and aggregated locally, before it gets a chance to accumulate. Then, partial results instead of full data are sent to the Cloud for stream processing. Batch processing is still used to complement this online dimension with a machine/deep learning dimension and gain more insights based on historical data (e.g. discover new correlations and patterns). The ultimate goal is to have an online/real-time front-end for processing on the Edge, close to where data streams are generated, while the Cloud will only be used for off-line back-end processing, mainly dealing with archival, fault tolerance and also further processing that is not time-critical. This hybrid approach enables edge analytics to detect *what* is happening with a monitored object, while Cloud analytics allows to understand *why* this is happening.

However, leveraging this dual approach in practice raises some significant challenges mainly due to the way in which stream processing engines organize the analytics workflow. Both Edge and Cloud engines create a *dataflow graph of operators* that are deployed on the distributed resources; they devise an execution plan by traversing this graph. In order to execute a request over such hybrid deployment, one needs a specific plan for the edge engines (e.g. Edgent), another one for the cloud SPEs (e.g. Spark) and to ensure the right interconnection between them thanks to an ingestion system (e.g. Kafka). Hence, with this non-uniform approach, the burden of connecting systems together and dividing the computation between them is left to users on a per application basis. Moreover, manually and empirically deploying this analytics pipeline (Edge-Ingestion-Cloud) can lead to sub-optimal computation placement with respect to the network

cost (i.e., high latency, low throughput) between the Edge and the Cloud.

In this paper, we argue that a *uniform approach* is needed to bridge the gap between Cloud SPEs and Edge analytics frameworks in order to leverage a single, transparent execution plan for stream processing in both environments. We introduce **Planner**, a streaming middleware capable of finding *cost-efficient cuts of execution plans between Edge and Cloud*. Our goal is to find a distributed placement of operators on Edge and Cloud nodes to minimize the stream processing makespan. This uniform view of the whole streaming pipeline is particularly novel and allows the placement to meet the constraints of throughput capacity of the Edge and Cloud resources as well as the bandwidth and latency limits of the network.

The contributions of this work are summarized as follows:

- We introduce a **resources model** for stream processing and **a cost model** for the streams flowing from an Edge operator to a Cloud-based one (Section IV);
- We **formulate the problem of operator placement** for an execution graph across distributed Edge and Cloud resources, with the objective of minimizing the network cost and the makespan (Section IV-A);
- We present **Planner**, a streaming middleware capable of automatically deploying fractions of the computations across Edge and Cloud, as a proof of concept (Section V);
- We perform comprehensive **real-world micro-benchmarks** showing that Planner reduces the network usage by 40% and the makespan by 15% compared to state-of-the-art. (Section VI).

II. Context and motivation

This section provides the background for our work and introduces the problem statement.

A. Infrastructure

A common infrastructure for stream processing is split into two layers: the Edge, hosting the devices which generate data, and the Cloud, used for ingestion (i.e., gathering data from devices and aggregating it into streams) and processing (i.e., the analytics on the incoming streams). While the Edge devices are becoming more and more resourceful and energy-efficient, the Cloud has (order of magnitude) more computing power. In this paper, we assume that the Cloud has enough resources to process the whole dataset after data collection. What remains prohibitive, however, is the high latency of the wide area networks connecting the Edge and the Cloud, leading to significant network usage and costs, and increased end-to-end processing latencies.

B. Data streaming

A data *stream* is an unbounded collection of atomic items. Processing *operators* consume streams and produce values (e.g., reduce, aggregate) or new streams (e.g., map, filter) using some User Defined Functions (UDFs). For instance, a *map* operator can transform a stream of temperature measurements into a stream of heat wave alerts). An operator source only produces streams (e.g reads items from files, devices) and, therefore, it irrigates the computation pipeline with data. An operator data sink only consumes data (e.g, writing data in a file).

Operators can be split into two categories: *stateless* and *stateful*. A stateless operator (e.g., map) processes items independently, one at a time, and consequently doesn't need to save its "state" in case of failures. In contrast, a stateful operator (e.g., reduce) processes items according to its local state (e.g., a rolling sum) or aggregates items and processes them by bucket (e.g., windows [14]).

C. Stream processing graphs

A common abstraction for modeling stream computations are the the *stream graphs*. They are directed acyclic graphs composed of operators (the vertices) interconnected by data streams (the edges).

We refine the notion of stream graph into a *weighted* DAG in order to model the network usage induced by streams and their sources (i.e, the average rate of events flowing through a stream). More formally $G_{st} = (V_{st}, E_{st}, \mathscr{W}_{st})$ denotes a stream graph where V_{st} is the set of operators, E_{st} is the set of streams and $\mathscr{W}_{st} : E_{st} \cup Sources \longrightarrow \mathbb{R}^+$ is the network usage. An operator o is composed of an UDF f_o and of a type τ_o (e.g., map, reduce) that describes an input and an output contract [16]. An input contract describes how the input items are organized into subsets that can be processed independently (e.g, by parallel instances) whereas an output contract denotes additional semantics information of the UDF (e.g, indicates that a UDF is stateless). The output contract can also give bounds for the selectivity s_o of an operator o. The selectivity [25] is the ratio of the output items rate over the input one of an operator (e.g, an operator issuing two items for one input has a selectivity of 2). For the sake of clarity, we summarize operator characteristics in Table I.

D. Problem statement

In order to run a computation, one needs to deploy the stream graph on the underlying infrastructure, i.e, to place operators on nodes. This mapping is called the *execution plan*. SPEs today can do such schedules either for the Cloud or for the Edge, *separately* (e.g., Spark deploys its execution plan on the Cloud, Minifi on the Edge).

In the case of complex hybrid infrastructures mixing both Edge and Cloud, however, the burden to define the partial computations, i.e., subgraphs, to be executed on each infrastructure, is delegated to the user. In many cases, this may lead to sub-optimal performance.

III. Models

In this section, we present the abstractions we leverage to model the resource graph on which the stream computation relies, as well as the network cost model that our approach aims to minimize. In Table II we summarize the notations used throughout the paper.

Type	Selectivy s_o	Locally-replicable $\mathscr{R}o$	Combination a_{τ_o}
Map	1	1	Id
FlatMap	≥ 0	1	Id
Filter	≤ 1	1	Id
Split	1	1	Id
Select	1	1	Id
Fold	1	0	Id
Reduce	1	0	Id
Union	1	1	$\sum$
Connect	1	0	min
Window	parametric[a]	0	Id

TABLE I: Operators overview. *Map:* takes one item and produces one item. *FlatMap:* takes one item and produces zero, one, or more items. *Filter:* takes one item and produces zero or one items. *Split:* splits a stream into two or more streams. *Select:* selects one or more streams from a split stream. *Fold:* combines the current item with the last folded value and emits the new value. *Reduce:* combines the current item with the last reduced value and emits the new value with an initial value. *Union:* union of two or more data streams. *Connect:* connects two data streams retaining their types. *Windows:* groups the data according to some characteristic.

A. Resources model

The computing and network resources used to execute a computation can be represented as a directed graph:

$$G_{res} = (Devices \cup \varsigma, E_{res}, \mathscr{W}_{res}) \tag{1}$$

where $Devices = \{d_i\}_{i \leq f}$ represent the set of Edge devices and ς represent the Cloud computing units. We aggregate the Cloud nodes in one logical node ς since we consider the Cloud powerful enough to run the full workflow (after collecting the data) and because we delegate the inner placement to the SPE (i.e., an engine like Spark or Flink will map the subgraph identified by our approach for cloud execution on the actual cloud nodes). E_{res} denotes the physical links between Edge devices and Cloud such as $Devices \times \varsigma \subseteq E_{res}$. We do not model other links[b] since we focus on the bottlenecks of the network between Edge and Cloud. Finally, $\mathscr{W}_{res} : E_{res} \to \mathbb{R}^+$ represents the cost of the transmission of an item through a link. We use a per item approach since the size of an item can arbitrary vary according to the nature of the UDF and there is no generic introspection mechanism to distinguish the shape of an item (e.g, items are arbitrary Java objects in Flink).

Special care should be taken when modeling sources, as they can produce data coming from multiple physical nodes. For instance, let us take some workflow monitoring crops where a data source aggregates (thanks to an ingestion queue) the temperature coming from several connected thermometers in order to increase reliability. We model such data dependencies by defining for each source $s \in Sources$ the group $g(s)$ of Edge devices that host the raw data used by s. Reciprocally, we define $g^{-1}(u)$ the group of sources using raw data hosted in the node u.

[b]Namely, we do not assume any hypothesis for the network between edge devices.

Symbol	Description
G_{st}	Stream graph representing the application
V_{st}	Set of vertices (operators) of G_{st}
V_{st}^d	Set of vertices run by the device $d \in Devices$
E_{st}	Set of edges (streams) of G_{st}
$\mathscr{W}_{st}(i,j)$	Communication usage induced by the stream $(i,j) \in E_{st}$
f_o	UDF executed by operator $o \in V_{st}$
τ_o	Type of operator $o \in V_{st}$
s_o	Selectivity of $o \in V_{st}$
G_{res}	Graph representing computing and network resources
$Devices$	Subset of vertices (computing Edge devices) of G_{res}
E_{res}	Set of edges (links between Edge and Cloud nodes) of G_{res}
$\mathscr{W}_{res}$	Transmission cost of an item through a link $l \in E_{res}$
d_i	An Edge device $d_i \in Devices$
ς	Represents all the cloud nodes
$g(o)$	Set of nodes ($g(o) \subset Devices$) that hold part of the raw data used by $o \in Sources$
$g^{-1}(u)$	Set of sources ($g^{-1}(u) \subset Sources$) that used part of the raw data of $u \in Devices \cup \{\varsigma\}$
a_τ	Aggregation function for operator of type τ
$\mathscr{C}_l^s$	Communication cost induced by the stream sE_{st} on link $l \in E_{res}$
C_{d_1}	Placement constraint for Edge device $d_i \in Devices$
$\mathscr{P}$	Operator placement
$\mathscr{T}$	Trace of items
$\mathscr{R}$	Local-replication indicator function

TABLE II: Main notations.

With this resource model, we can use different cost functions depending on the metric we want to optimize (e.g, we can use an energetic cost per item or the latency of the link).

B. Network cost model

Due to the black-box nature of UDFs, we approximate the network usage of the streams over the links between Edge and Cloud. The network usage of a stream $(i,j) \in E_{st}$ depends on the input rate of operator i, the selectivity of i and its type of τ_i. This is formally expressed as follows:

$$\mathscr{W}_{st}(i,j) = \begin{cases} s_i * a_{\tau_i}\left(\mathscr{W}_{st}(\xi_{i_1}), \ldots, \mathscr{W}_{st}(\xi_{i_l})\right) & \text{if } i \notin Sources \\ \mathscr{W}_{st}(i) & \text{otherwise} \end{cases} \tag{2}$$

where $(\xi_{i_1}, \ldots, \xi_{i_l})$ are the input streams of i in G_{st}, $a_{\tau_i} : \mathbb{R}^k \to \mathbb{R}$ is the weight aggregation function for an operator of type τ_i and where k is the input arity of i. a_{τ_i} describes the combination of network usage of incoming streams (cf.

Table I). Furthermore, $\mathscr{W}_{st}(i)$ denotes the average event rate produced by a source i, which should be estimated using static analysis or by probing the source at runtime. Finally, $\mathscr{C}_{\xi}^{l} = \mathscr{W}_{res}(l) * \mathscr{W}_{st}(\xi)$ denotes the communication cost of a stream $\xi \in E_{st}$ flowing through a link $l \in E_{res}$.

IV. Uniform Stream Graph Placement

Our key idea for finding the ideal cut (between Cloud and Edge) of the stream graph is to solve an optimisation problem for placement, while trying to minimise the network cost. We formulate this problem and its optimisations in this section.

A. The placement problem

Not all operators can be executed on Edge devices due to their limited computing power, memory, battery life or simply because some operators are not supported by the Edge analytics frameworks. Therefore, for each Edge device d we encode such a restriction in a constraint C_d. A stream graph H can be placed to the device d if and only if it satisfies the constraint C_d, denoted by $H \models C_d$.

The placement problem aims at minimizing the global communication cost (and, consequently, the stream processing makespan) by executing some computations on the Edge devices instead of moving all the data to the Cloud. This is formally expressed as follows:

$$\min_{\mathscr{P}} \sum_{\{V_{st}^d\} \in \mathscr{P}} \sum_{\xi \in E_{st} \cap (V_{st}^d \times (V_{st} \setminus V_{st}^d))} \mathscr{C}_{\xi}^{(d,\varsigma)} \tag{3}$$

subject to:

$$V_{st}^d \models C_d$$

where ξ is a stream flowing over a link between Edge and Cloud and $\mathscr{P}$ denotes the set of operators that should be executed on each Edge device. A placement $\mathscr{P}$ is defined as follows:

$$\mathscr{P} = \bigcup_{d \in Devices} \{V_{st}^d\} \tag{4}$$

where V_{st}^d is the subgraph of G_{st} mapped to the device d. The remaining part of the workflow will be executed in the Cloud.

We can define a placement problem as a conjunction of independent placement problems for each device by restricting the constraint placement C_d for each device d. A *local placement problem* for a device d is then formally stated as follows:

$$\min_{V_{st}^d} \sum_{\xi \in E_{st} \cap (V_{st}^d \times (V_{st} \setminus V_{st}^d))} \mathscr{C}_{\xi}^{(d,\varsigma)} \tag{5}$$

subject to:

$$V_{st}^d \models \tilde{C}_d$$

where the restricted constraint is as follows:

$$\tilde{C}_d = C_d \wedge Source(X) \subset g^{-1}(d) \tag{6}$$

$Source(X)$ is the set of sources of the candidate subgraph and C_d is the previous constraint for the device d.

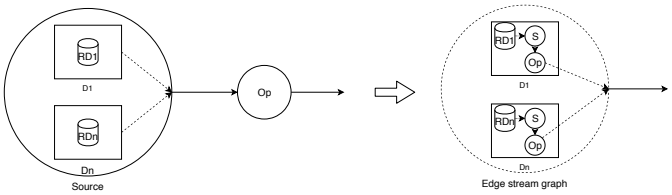

Fig. 1: Raw-data locally-aware optimization where Op is an operator, D_i denotes an edge device and RD_i is the raw data (hosted by D_i) used by the source.

B. Locality-awareness optimization

We introduce a locality-aware optimization in order to address the *local placement problem*. This optimization aims at deploying operators near the raw data in the Edge by allocating one version of the operator per device and to further collect the results in one stream inside the ingestion system (Figure 1).

An operator o is *locally-replicable* (and denoted by the indicator function $\mathscr{R} : V_{st} \longrightarrow \{0,1\}$) if and only if the former optimization applied to o preserves the equivalence of computation defined using the following equivalence relation.

Computation equivalence. In order to compare computations done by two stream graphs composed of deterministic[c] operators, we define a notion of equivalence based on the outputted items. We state that two stream graphs G_1 and G_2 are equivalent if for any trace $\mathscr{T}$ (an ordered sequence of input items for each input streams) $G_1(\mathscr{T}) == G_2(\mathscr{T})$, where $G_1(\mathscr{T})$ denotes the content of the output streams produced by G_1 when applied to $\mathscr{T}$. Two streams are equal if they coincide for any finite prefix.

Theorem 1. *A stateless deterministic operator is locally-replicable.*

The idea of the proof for operator o is as follows. Let us split the trace in sub-traces: $\mathscr{T} = \bigcup_i \mathscr{T}_i$ (one for each device involved). Now, by combining the local results with the same interleaving $\bigcup_i o(\mathscr{T}_i)$ we obtain $o(\mathscr{T})$ since the output of o only depends of the current input item (because o is stateless and deterministic). An overview of locally-replicable operators is available in Table I.

Nothing can be said for stateful operators due to the unknown behaviour of the UDFs. Indeed, a stateful operator can be locally-replicable (e.g, the identity map can be simulated with a reduce operator by ignoring its state). In turn, we can exhibit the following situation where an operator is not locally-replicable. Let us take two devices one providing the odd numbers and the other the even ones, and a source which outputs the data for both devices. Eventually, the produced stream is consumed by a reduce operator computing a rolling sum (i.e, the sum of the current item with the last reduced value). If we take a trace that alternates even and odd value,

[c]The notion of equivalence is not defined in the non-deterministic case. Some weak-equivalence can be defined by defining $G_1(\mathscr{T})$ as the set of possible output traces.

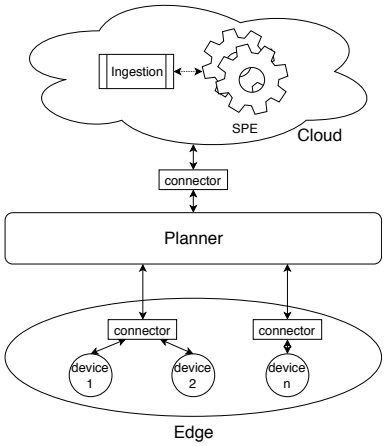

Fig. 2: A hybrid infrastructure where Planner is used to deploy parts of the computation on edge devices. The connectors are small pieces of software used to plug Planner with other cloud-based analytics systems (e.g., Apache Flink).

then before the optimization the output stream is composed of odd numbers and after it is composed of even numbers.

V. Planner Overview

We implement this approach into a proof of concept, called Planner - a streaming middleware unifying Edge and Cloud analytics. Planner automatically and transparently delegates a light part of the computation to Edge devices (e.g., running embedded edge processing engines) in order to minimize the network cost and the end-to-end processing time of the Cloud based stream processing. It does so as a thin extension of a traditional cloud-based SPE (e.g., Apache Flink in our case) to support hybrid deployments, as shown in Figure 2.

In this section, we first introduce the design principles backing such an approach, then we provide an architectural overview of Planner. We particularly zoom on its scheduler, which is responsible for finding cost-efficient cuts of execution plans between Edge and Cloud.

A. Design principles

Planner has been designed according to the following three principles:

1) A transparent top-down approach: Streaming applications are submitted by users unchanged to the SPEs. The latter translate them into streaming graphs (execution plans) that Planner intercepts and divides between Cloud and Edge. Therefore, Planner is fully transparent for users. As a side effect, this top-down approach is well suited especially for Cloud plans, which tend to be more expressive and complex than the Edge ones, as they leverage many global stateful operators (e.g., window based operators).

2) Support for semantic homogeneity: Edge devices are considered to be homogeneous in terms of semantics of the computation, i.e., each device provides data to the same group of sources ($\exists A \subset Sources, \forall d \in Devices, g^{-1}(d) = A$). This restriction is a drawback of the transparency. Indeed, the graphs submitted to Planner are limited by the expressivity of the SPE. And most SPEs are not designed to track the provenance of the data produced by sources. However, this limitation does not apply to the former resource model which supports heterogeneous types of devices.

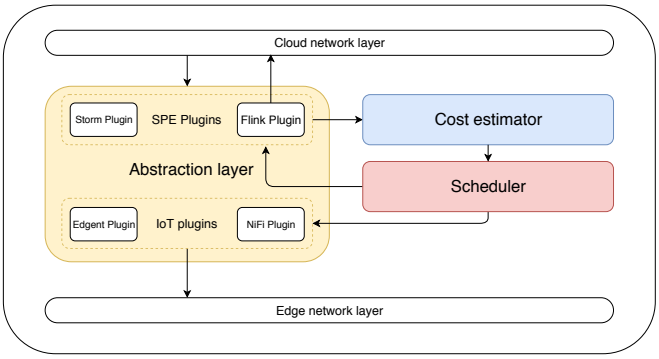

Fig. 3: The Planner architecture.

3) Support for interaction with SPEs: Planner is system agnostic: the core of the middleware is not aware of the details of the cloud or edge SPE but only of some abstract representation. This allows any backend to be easily plugged to Planner thanks to specific external connectors. A connector is responsible of the deep interaction with external systems (e.g exporting plans to Planner and importing and executing instructions from Planner). Additionally, the richer model of Planner can be used to improve the computation made by the SPE by permuting operators in order to reduce the overall network usage.

B. Architecture overview

Planner takes as input a stream graph expressed in the "dialect" of an SPE and outputs multiple stream graphs: one for the Cloud SPE and one per (groups of) Edge device(s). To enable this behavior, Planner is structured in three layers (Figure 3): an Abstraction Layer, a Cost Estimator, and a Scheduler.

1) The Abstraction Layer: translates input stream graph (coming from the cloud SPE connectors) to an abstract graph which is an instance of the one presented in Section II-C. However, some relevant information usually lacks from the cloud SPE streaming plan (e.g. operator selectivity is not present in Apache Flink plan). In such a case, we provide default values based on the operator type (using Table I). Conversely, this layer also translate an inner abstract representation to cloud or edge SPE "dialects".

2) The Cost Estimator: enhances the abstract stream graph with the network usage of the streams. This is computed by applying the model represented in the Equation 2 in a topological order starting from the sources. It is also in charge of maintaining the transmission costs of the links between Edge and Cloud (in the current implementation we assume a constant transmission cost).

3) The Scheduler: selects which part of the stream graph should be deployed to the Edge in order to minimize the

overall cost of the placement by applying the optimization presented in Section IV-B. Moreover, it restricts the placement constraints (see Equation 5) in order to process each Edge device independently.

Algorithm 1 Placement algorithm for an Edge device d

Require: $S = g^{-1}(d)$
1: $Current \leftarrow S$
2: $Opened \leftarrow \bigcup_{s \in S} \{x \in Succ(s) \mid Pred(x) \subset Current\}$
3: $Closed \leftarrow S$
4: **while** $Opened \neq \varnothing$ **do**
5: Pick u in $Opened$
6: Let H the subgraph of G_{st} induced by $Current \cup \{u\}$
7: **if** $H \models \tilde{C}_d$ **then**
8: $Current \leftarrow Current \cup \{u\}$
9: $Applicants \leftarrow \{x \in Succ(u) \mid Pred(x) \subset Current\} \setminus Sinks$
10: $Opened \leftarrow Opened \cup Applicants \setminus Closed$
11: **end if**
12: $Opened \leftarrow Opened \setminus \{u\}$
13: $Closed \leftarrow Closed \cup \{u\}$
14: **end while**
15: $Border \leftarrow N(Current) \setminus Current$
16: F is the subgraph of G_{st} induced by $Border \cup Current$
17: **return** a minimum (S, $Border$)-cut in F

We use a two-phase approach, as shown in Algorithm 1. Firstly (lines 4-14), we do a traversal of the graph and extract the maximal subgraph that can be placed on an Edge device with respect to the constraint satisfaction. $Current$ denotes the set of operators that will be moved to the device d. Note that if a data source has no successor in $Current$, then it will remain in the Cloud. Moreover, $Current$ verifies:

$$Pred(Current) \subset Current \wedge H \models \tilde{C}_d \wedge Current \cap Sinks = \varnothing \quad (7)$$

where $Pred(o)$ is the set of the predecessors of o in G_{st} and H is the subgraph of G_{st} induced by $Current$. $Opened$ denotes the operators to process such that $\forall x \in Opened, Pred(x) \subset Current$. $Closed$ denotes the operators that have been processed.

Secondly, we compute a minimum (S, $Border$)-cut (using the StoerWagner algorithm [32]) where $Border$ denotes the external neighbours of $Current$ operators. In the implementation (unlike the model) we do not limit the computing power of an edge device [d] since refining the resources needed to run an operator would require to analyze arbitrary UDFs using static analysis or online profiling.

Let us discuss the optimality of the previous algorithm with respect to the *local placement problem* (see Equation 5), depending on the nature of the constraint. If $\tilde{C}_d$ does not encode any notion of capacity of the Edge devices (for instance the constraint could be: "the operator is locally-replicable" or "the Edge SPE cannot run this kind of computation") then this algorithm gives an optimal placement by definition of min-cut. Otherwise, if some capacity constraint is encoded in the constraint (e.g., maximun memory consumption), there is an underlying knapsack problem. One way to improve this algorithm is to refine the selection of u (line 4).

[d] We assume that an edge device can compute the subgraph of operators that Planner sends to it; in practice this graph is small without complex operators.

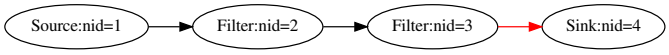

Fig. 4: The source produces taxi ride items then they are filtered in order to get statistics on the rides in New York city (e.g., rides that have not started, rides taking too much time) and eventually stored in a Kafka topic.

The complexity $T(n)$ of the algorithm does not depend on the number of Edge devices and it is defined as $O(n^2 \log n + n\alpha(n) + nm)$ where n is the number of operators, m is the number of streams and $\alpha(n)$ denotes the complexity of the constraint satisfaction[e]. In practice, the number of operators is small (from tens to a few hundreds) and stream graphs are commonly sparse because most of the operators have an input (and output) arity of one or two.

In order to process all the devices, we simply apply the former algorithm for each device. This naive approach leads to a complexity of $O(|Devices| * T(n))$. Therefore, it is linear on the number of devices (which can reach several millions). However, we can refine the previous algorithm, in order to scale more, by grouping devices and by applying the Algorithm 1 for each group. Grouping should be done according to device nature, i.e, a group of connected cars and a group of connected thermometers. More formally, a group is the set of devices that share the same $g^{-1}(d)$). Finally, if the characteristics (e.g., computer power) of the devices in the same group are not close enough, we can use a hierarchical structure. We thus create subgroups based on characteristic similarities; the input graph of a subgroup is the output stream graph of its parent group. Eventually, Planner can be decentralized by spanning an instance of Planner per bunch of (groups of) devices and each of this instance have a copy of the full stream graph. This can be done since Planner takes advantage of the local placement problem (Eq. 5) where (group of) devices can be processed independently.

VI. Validation

A. Experimental setup

We emulate a hybrid Cloud and Edge platform on the Grid'5000 testbed. Cloud nodes are located on the *paravance* cluster in Rennes and Edge devices on the *graphene* cluster in Nancy. Cloud nodes are composed of 2 x E5-2630v3 (8 cores/CPU) with 128 GB of memory and they are connected to the network by 2 x 10 Gbps links. Edge devices run on one core of an Intel Xeon X3440 with 16 GB of memory and

[e] We can obtain $\alpha(n) = O(n)$ with simple constraints expressing computing power or memory limitations.

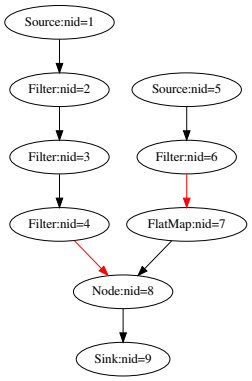

Fig. 5: The source 1 produces taxi ride items and the source 5 taxi fare ones. This workflow computes the set of well-formed night rides in NY city and each ride is joined with its fare (by operator 8).

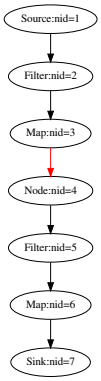

Fig. 6: This benchmark calculates every 5 minutes popular areas where many taxis arrived or departed in the last 15 minutes.

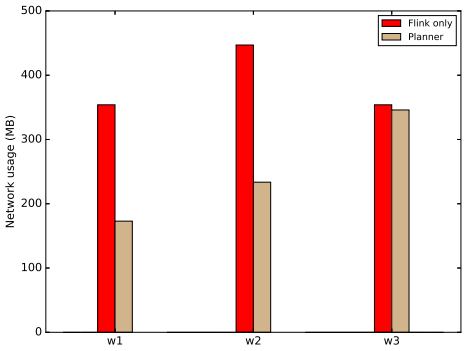

Fig. 7: Network usage over links between Edge and Cloud induced by application execution where $w1$ denotes the workflow in Figure 4, $w2$ denotes the workflow in Figure 5 and $w3$ the one of Figure 6. The red bar corresponds to the whole computation in Cloud and the tan one corresponds to the usage of Planner.

connected to network by 1 Gbps links. For our experiments we have used up to 10 nodes and a total of 80 cores.

To emulate the high latency WANs connecting Edge and Clouds we use *tc* [6] to vary latency (with *netem*) and the available bandwidth (with *tbf*) until reaching the desired quality of service. Edge nodes on steroids (i.e., the least powerful nodes in the Grid'5000, yet quite powerful for an average Edge device) should not impact the validation since the network is constrained with *tc* and we ignore performance capacity during placement (we only target expressiveness constraints).

For the validation we use Apache Flink as the Cloud SPE and Apache Edgent as the Edge SPE. We have chosen this Edgent-Flink duo for simplicity of integration since both are written in Java and are based on arbitrary UDF-operators. We use one cloud node for Flink (with one *jobmanager* and one *taskmanager*), one cloud node for the ingestion system (one Apache Kafka broker). We deploy five Edgent devices on distinct Edge nodes and we collocate Planner with the *jobmanager*. Interconnection with Planner is done via a dedicated Flink connector collocated with the *jobmanager* and five Edgent connectors hosted in each Edge node.

B. Experimental protocol

We ran two real-life application dataflows presented in Figure 4 and Figure 5 where the red arrows represent the cut found by applying Algorithm 1. They rely on the [f] of the New York City Taxi dataset [21] composed of data containing fares (1.5M entries) and rides (3M entries) description for 15K distinct taxis, with a total size of about 10GB. The rides dataset contains especially the start location, the stop location, the start time, the end time and the fares dataset contains in particular the tip, the toll and the total fare of a ride. For each dataflow, we compare two deployment scenarios where raw data are hosted in the Edge devices. For the first one, the whole computation is processed on Cloud with Flink (with data served from the Edge using Kafka). For the other one, Planner is used to deploy part of the computation to the Edge devices.

[f] http://training.data-artisans.com/exercises/taxiData.html

C. Results

In our first series of experiments, we measured the reduction in terms of transferred data with our approach. As seen in Figure 7, Planner is able to reduce the network usage over links by 51% for the workflow $w1$ (Figure 4) and by 43% for the workflow $w2$ (Figure 5). Our Cost Estimator is mainly based on the selectivity and filter operators have the lowest selectivity (sinks excepted). Therefore, the scheduler will place as much filter operators as possible on the Edge. However, even for one of the worst cases for Planner (the workflow $w3$ in Figure 6, where there is global stateful operator - here a time window - near the sources and a very light preprocessing - here a light clean of data), our approach is still able to reduce the network usage compared to vanilla Flink.

In the second series of experiments, we measured the reduction of the end-to-end processing latency (Figure 8) and of the makespan (Figure 9) with our approach when the bandwidth between Edge devices and Cloud varies. As seen in both plots, Planner gains over vanilla Flink (all execution in the Cloud) is smaller than for the network usage because of the lack of inner optimizations in Edgent. For instance in Flink, operators are grouped in the same logical operator (and then in the same thread) in order to optimize computation. Moreover, we can observe that the gain brought by Planner is better for $w2$ than $w1$ for the latency and, inversely, better for $w1$ than $w2$ for the makespan. This is due the fact that there is a connect operator linking fares and rides according to the *taxi id* furthermore it also explains the outlying results (standard deviation) in Fig. 8. Minimizing the network usage between Edge and Cloud

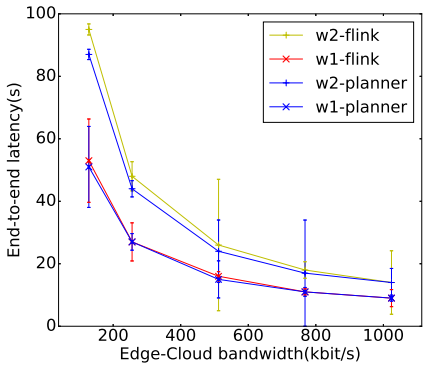

Fig. 8: End-to-end processing latency: the green and red lines correspond to the whole computation in the Cloud and the blue one corresponds to the Planner approach.

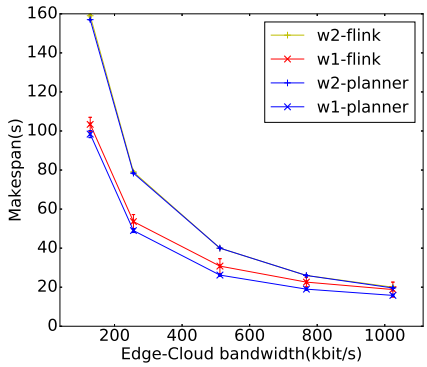

Fig. 9: Makespan: the green and red line correspond to the whole computation in the Cloud and the blue one corresponds to the Planner approach.

reduces the congestion of this operator and, consequently, single rides (or fare) stalls less and eventually the latency gain is greater. Conversely, Planner performs better on $w1$ than $w2$ for the makespan because the network usage decreases more for the first one. Overall, we notice an average of 15% improvement for the makespan and the latency, proportional with the bandwidth between the Edge and Cloud.

VII. Discussion

A. Assumptions

1) Communication bottleneck: We focus on the network as the main bottleneck of stream processing over Edge and Cloud systems, starting from the time skew between the generation time and the processing time of events [14]. Nevertheless, we also consider other limitations of the overall performance (e.g., power, memory, computing of an Edge device). This is the role of the constraints introduced in Section IV-A.

2) Cloud power: We consider that systems deployed on the Cloud are able to run the whole computation. For common use cases, this assumption holds. However, there are some situations where one datacenter can not run the whole process (e.g. more than 80 datacenters are involved in the processing of the data produced by MonALISA [28] monitoring ALICE [12] one of the four experiments of the LHC). In this case, one can use a geo-distributed cloud middleware for stream processing (like SpanEdge [30]) as a cloud SPE for Planner.

3) Additional metrics: Planner has been tailored to focus on the optimization of network usage and, thus contributing to the makespan reduction.. However, it can also optimize other metrics (e.g. the throughput of the application) by careful selection of a convenient definition of the transmission cost $\mathscr{W}_{res}$ (defined in Section III-A) for an item.

B. How to overcome limitations

1) Homogeneity of IoT devices: Each IoT device, independently of its intrinsic characteristics, will run the same subgraph - with distinct datasets) due to the *semantic homogeneity* (see Section V-A2). This can be mitigated by letting the user add optional annotations to the cloud SPE workflow in order to specify which group (i.e., $g(o)$) of devices provides raw data to a source operator. Moreover, annotations can also be used to distinguish non Edge sources (e.g., a DB connector). Finally, the transparency can be preserved with some SPEs by encoding annotations in operator names or uids (if available).

2) Graph optimization: Planner does not do any execution plan optimization (e.g. [26] [25]), contrary to Apache Flink for instance. A plan optimization is a rewriting of the stream graph in order to improve its shape for further processing (for instance, by pushing filter operators near the sources). Moreover, stream graph rewriting is mainly a matter of finding operators that commute using static analysis (distinct for each SPE). Therefore, the static analysis should be done before the Abstraction Layer and the rewriting should be combined with the Cost Estimator.

3) Increased model accuracy: One point of improvement is the accuracy of the cost model and particularly the measure of operators metrics (e.g., selectivity or record rate of sources). This is due to the unknown behaviour of operators. However, we can enhance the accuracy of the record rate of a source by refining the measurements at runtime with an embedded probe. The same approach can be applied in order to improve the accuracy of the selectivity of operators.

C. Take-aways

1) What Planner is: Planner is a streaming middleware capable of automatically and transparently deploying parts of the computation across Edge and Cloud resources. Furthermore, it is modular enough to be plugged with any combination of Cloud and Edge SPEs via external connectors.

2) On the generality of the Planner approach: The combination of Cloud and Edge SPEs on which Planner works is only limited by their expressiveness. For instance, an UDF-based Cloud SPE cannot be plugged with a non UDF Edge SPE since the framework will not be able to run the exported operators. But slight discrepancy of expressiveness (e.g., some small subset of non-supported operators) can be tolerated thanks to the constraints introduced in Section IV-A.

3) What Planner is not: Planner does not deploy streaming systems on Edge device or in the Cloud (as it expects the SPEs to be already deployed) and it is neither intended to place operators at the granularity of physical nodes (this is delegated to the SPEs, which schedule the received sub-graphs). Besides, Planner is not yet preserving consistency. However, with the right combination of Cloud SPEs, Edge SPEs and ingestion systems, some levels of consistency could be guaranteed (for instance, *exactly once* using Kafka and Flink). Also, Planner does not ensure any data persistence. Indeed, by moving computation to Edge devices parts of the data will never reach the ingestion system and therefore will not be persistently stored.

VIII. Related Work

We divide the state-of-the-art into three categories: 1) classical stream processing systems, 2) systems for hybrid processing and 3) works on general stream processing optimizations.

A. Stream Processing Engines

1) Edge analytics frameworks: Edge analytics frameworks, like Apache Edgent [4], Apache MiNiFi [5], are used to execute a data stream application on an Edge device. They are optimized to do local light-weight stream processing. Such frameworks commonly export results to an ingestion system, like Apache Kafka [27] or RabbitMQ [33].

2) Cloud SPEs: A common approach to execute a stream graph is to use SPEs that will take care of the deployment of the computation to many physical nodes, their management and their fault-tolerance. Several SPEs exists (e.g, Apache Flink [19], Apache Spark [34], Amazon Kinesis [7], Google Dataflow [8]). They are mainly designed and optimized in order to run in the Cloud and particularly in a single data-center [30].

B. Hybrid approaches

A new paradigm has emerged which combines Cloud-based and Edge analytics in order to do real-time processing at the Edge (for some timely but inaccurate results) and offline processing in Cloud (for late but accurate results) inside the same application.

Several companies are providing solutions (e.g, Azure Stream [9], IBM Watson IoT [10], Cisco Kinetic [11]) that should ease the deployment of the stream processing on Edge devices and to interconnect with their own Cloud-oriented SPEs. However, they are provided "as a-service" and the user is dependent of the companies' platforms.

SpanEdge [30] focuses on unifying stream processing over geo-distributed data-centers (between a main datacenter and several near-the-edge ones) in order to take advantage of user and data locality to reduce the network cost and the latency. They are placing computations on distinct data-centers whereas we are targeting locality-aware placement on edge devices.

The authors of [23] also use a weighted stream graph but they try to find the optimal solution whereas Planner uses an heuristic placement that can be efficiently done (see V-B3).

Echo [31] generalizes data stream processing on top of an Edge environment. It maps operators onto Cloud nodes and Edge devices in order to take advantage of the unused computing power available in the Edge. Unlike Echo, we are using a locality-aware placement approach in order to minimize communication cost. Furthermore, the placement of Echo works at the granularity of nodes (e.g., it bypasses the placement strategies of the SPEs and their potential optimisations) whereas Planner places stream sub-graphs to systems, (leveraging the SPEs strategies to place operators onto nodes and benefit of their inner optimizations).

C. Optimizing stream processing

There are two orthogonal approaches (commonly applied in sequence [15]) that try to optimize the processing of a stream graph: graph rewriting and graph scheduling.

1) Graph rewriting: rewrites the input plan in an equivalent one that should improve the performance of the computation (e.g. [26], [25]). This job is mainly done by permuting operators and by replicating them in order to improve the parallelism.

2) Placement optimization: This focuses on the mapping of the operators to physical nodes. A lot of modeling and algorithmic work has been accomplished in the context of [29] [23] [20] [17]. This kind of optimizations are not performed (and neither intended to be) by Planner, which in turn delegates fine grained placement to other systems.

IX. Conclusion

In this paper, we address a challenging problem for *hybrid stream processing* on infrastructures combining Cloud and Edge components in a shared system. With this paradigm, computation placement is usually done manually. Besides being a burden for users this can lead to sub-optimal computation placement with respect to network cost between the Edge and the Cloud.

We argue for a uniform approach in order to leverage a single, transparent and automatic execution plan on such a hybrid platform. We provide a model of a hybrid infrastructure and a generic model of the network cost over Edge and Cloud links. From them, we define a plan placement problem in order to minimize the makespan and the network cost. We restrict this placement into a local one which processes (groups of) agents independently in order to improve scalability. Then we introduce a new raw-data locality-aware optimization which preserves the semantics of the computation and we derive a scheduler. As a proof of concept we implement Planner, a streaming middleware that automatically partitions the execution plans across Edge and Cloud. We evaluate our work by setting up an hybrid architecture on Grid'5000, where we deploy Planner with Apache Flink and Apache Edgent. By running real-world micro-benchmarks, we show that Planner reduces the network usage by more than 40% and the makespan by 15%

As future work, we plan to add optional workflow annotations and then enable support for heterogeneous sources.

Moreover, we plan to introduce a new optimization (based on a weak equivalence of computation that guarantees not to introduce new behaviours) in order to export some stateful operators (e.g. reduce). Finally, we plan to switch from static placement to an adaptive one where metrics about operators (e.g. selectivity) and infrastructure (e.g. average throughput) are refined at runtime in order to increase the accuracy of the cost model and to periodically trigger the plan placement computation.

X. Acknowledgements

This work is supported by the ANR OverFlow project (ANR-15-CE25-0003).

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

References

[1] http://storm.apache.org/. [Online; accessed 9-August-2018].

[2] http://samza.apache.org/. [Online; accessed 9-August-2018].

[3] https://pulsar.incubator.apache.org/. [Online; accessed 9-August-2018].

[4] http://edgent.apache.org/. [Online; accessed 13-July-2018].

[5] https://nifi.apache.org/minifi/. [Online; accessed 13-July-2018].

[6] https://github.com/shemminger/iproute2. [Online; accessed 9-August-2018].

[7] https://aws.amazon.com/kinesis/. [Online; accessed 13-July-2018].

[8] https://cloud.google.com/dataflow/. [Online; accessed 13-July-2018].

[9] https://azure.microsoft.com/en-us/services/stream-analytics/. [Online; accessed 13-July-2018].

[10] https://www.ibm.com/internet-of-things. [Online; accessed 13-July-2018].

[11] https://www.cisco.com/c/fr_fr/solutions/internet-of-things/iot-kinetic.html. [Online; accessed 13-July-2018].

[12] Kenneth Aamodt, A Abrahantes Quintana, R Achenbach, S Acounis, D Adamová, C Adler, M Aggarwal, F Agnese, G Aglieri Rinella, Z Ahammed, et al. The alice experiment at the cern lhc. *Journal of Instrumentation*, 3(08):S08002, 2008.

[13] Daniel J Abadi, Don Carney, Ugur Çetintemel, Mitch Cherniack, Christian Convey, Sangdon Lee, Michael Stonebraker, Nesime Tatbul, and Stan Zdonik. Aurora: a new model and architecture for data stream management. *the VLDB Journal*, 12(2):120–139, 2003.

[14] Tyler Akidau, Robert Bradshaw, Craig Chambers, Slava Chernyak, Rafael J Fernández-Moctezuma, Reuven Lax, Sam McVeety, Daniel Mills, Frances Perry, Eric Schmidt, et al. The dataflow model: a practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. *Proceedings of the VLDB Endowment*, 8(12):1792–1803, 2015.

[15] Alexander Alexandrov, Rico Bergmann, Stephan Ewen, Johann-Christoph Freytag, Fabian Hueske, Arvid Heise, Odej Kao, Marcus Leich, Ulf Leser, Volker Markl, et al. The stratosphere platform for big data analytics. *The VLDB JournalThe International Journal on Very Large Data Bases*, 23(6):939–964, 2014.

[16] Dominic Battré, Stephan Ewen, Fabian Hueske, Odej Kao, Volker Markl, and Daniel Warneke. Nephele/pacts: a programming model and execution framework for web-scale analytical processing. In *Proceedings of the 1st ACM symposium on Cloud computing*, pages 119–130. ACM, 2010.

[17] Benjamin Billet and Valérie Issarny. From task graphs to concrete actions: a new task mapping algorithm for the future internet of things. In *MASS-11th IEEE International Conference on Mobile Ad hoc and Sensor Systems*, 2014.

[18] Flavio Bonomi, Rodolfo Milito, Jiang Zhu, and Sateesh Addepalli. Fog computing and its role in the internet of things. In *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*, pages 13–16. ACM, 2012.

[19] Paris Carbone, Asterios Katsifodimos, Stephan Ewen, Volker Markl, Seif Haridi, and Kostas Tzoumas. Apache flink: Stream and batch processing in a single engine. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 36(4), 2015.

[20] Valeria Cardellini, Vincenzo Grassi, Francesco Lo Presti, and Matteo Nardelli. Optimal operator replication and placement for distributed stream processing systems. *ACM SIGMETRICS Performance Evaluation Review*, 44(4):11–22, 2017.

[21] B Donovan and DB Work. New york city taxi trip data (2010–2013), 2014.

[22] Pedro Garcia Lopez, Alberto Montresor, Dick Epema, Anwitaman Datta, Teruo Higashino, Adriana Iamnitchi, Marinho Barcellos, Pascal Felber, and Etienne Riviere. Edge-centric computing: Vision and challenges. *ACM SIGCOMM Computer Communication Review*, 45(5):37–42, 2015.

[23] Rajrup Ghosh and Yogesh Simmhan. Distributed scheduling of event analytics across edge and cloud. *arXiv preprint arXiv:1608.01537*, 2016.

[24] Nithyashri Govindarajan, Yogesh Simmhan, Nitin Jamadagni, and Prasant Misra. Event processing across edge and the cloud for internet of things applications. In *Proceedings of the 20th International Conference on Management of Data*, pages 101–104. Computer Society of India, 2014.

[25] Martin Hirzel, Robert Soulé, Scott Schneider, Buğra Gedik, and Robert Grimm. A catalog of stream processing optimizations. *ACM Computing Surveys (CSUR)*, 46(4):46, 2014.

[26] Fabian Hueske, Mathias Peters, Matthias J Sax, Astrid Rheinländer, Rico Bergmann, Aljoscha Krettek, and Kostas Tzoumas. Opening the black boxes in data flow optimization. *Proceedings of the VLDB Endowment*, 5(11):1256–1267, 2012.

[27] Jay Kreps, Neha Narkhede, Jun Rao, et al. Kafka: A distributed messaging system for log processing. In *Proceedings of the NetDB*, pages 1–7, 2011.

[28] Iosif Legrand, C Cirstoiu, C Grigoras, R Voicu, M Toarta, C Dobre, and H Newman. Monalisa: An agent based, dynamic service system to monitor, control and optimize grid based applications. 2005.

[29] Peter Pietzuch, Jonathan Ledlie, Jeffrey Shneidman, Mema Roussopoulos, Matt Welsh, and Margo Seltzer. Network-aware operator placement for stream-processing systems. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*, pages 49–49. IEEE, 2006.

[30] Hooman Peiro Sajjad, Ken Danniswara, Ahmad Al-Shishtawy, and Vladimir Vlassov. Spanedge: Towards unifying stream processing over central and near-the-edge data centers. In *Edge Computing (SEC), IEEE/ACM Symposium on*, pages 168–178. IEEE, 2016.

[31] Sarthak Sharma, Prateeksha Varshney, and Yogesh Simmhan. Echo: An adaptive orchestration platform for hybrid dataflows across cloud and edge. In *Service-Oriented Computing: 15th International Conference, ICSOC 2017, Malaga, Spain, November 13–16, 2017, Proceedings*, volume 10601, page 395. Springer, 2017.

[32] Mechthild Stoer and Frank Wagner. A simple min-cut algorithm. *Journal of the ACM (JACM)*, 44(4):585–591, 1997.

[33] Alvaro Videla and Jason JW Williams. *RabbitMQ in action: distributed messaging for everyone*. Manning, 2012.

[34] Matei Zaharia, Reynold S Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J Franklin, et al. Apache spark: a unified engine for big data processing. *Communications of the ACM*, 59(11):56–65, 2016.